面向任务难度感知的检索行为研究*

■ 黄崑[1] 陈佳琦[1] 刘畅[2]
[1] 北京师范大学政府管理学院 北京 100875 [2] 北京大学信息管理系 北京 100871

摘 要：[目的/意义]以学习型搜索为背景，探索对于反映用户感知任务难度具有稳定、显著影响的交互行为指标。[方法/过程]依托课堂教学开展行为实验，以《信息技术应用》课程学生为研究对象，布置5项搜索任务，运用Morae软件进行过程录制，共收集39名学生的170条有效检索行为数据。根据录屏数据和任务难度感知评价数据，对不同难度下呈现显著差异的行为指标进行分析。结合课题组在2016年、2017年开展的两项实验数据，以及已有相关研究发现进行横向比较，归纳不同任务难度下行为指标呈现显著差异的稳定性。[结果/结论]最终获得具有26项行为指标的稳定行为指标体系，从探索式搜索过程及行为类型角度对用户在困难任务下的行为特点进行描述，并指出未来进一步研究的方向。

关键词：任务难度 检索行为 交互式信息检索 实验研究
分类号：G252.7
DOI：10.13266/j.issn.0252-3116.2022.03.010

1 绪论

探索任务多维属性特征与搜索交互过程可观测特征的联系，是近年来交互式信息检索领域的研究热点之一[1]。任务难度是任务的重要属性之一[2-3]，任务难度影响着检索交互行为、检索绩效以及检索体验[4]。在学习型搜索中，任务难度也会对搜索过程的学习效果、学习体验产生影响[5]。难度一般是指项目(如任务、题目等)的困难容易程度[6]。用户对搜索任务困难的感知可能来自搜索过程多方面的困难，与搜索交互过程中的用户、任务以及交互过程都有关系[7]，包括理解信息需求、开展搜索过程，以及解释结果相关性等困难[8]。为了提升用户的搜索绩效、搜索体验和学习收获，研究者们通过探索任务难度属性与检索行为的关系，力图构建基于检索行为的任务难度预测模型，为检索系统及时洞察用户的搜索意图、困境进行相应干预和支持提供参考[9-12]。检索行为与难度感知关系的稳定性影响着基于检索行为预测任务难度感知的效果。

在信息搜索任务难度方面，不少研究者已经开展了一系列工作。本研究在前期工作中开展了较为系统的综述性工作[13]，可以看到，不少研究从任务困难原因、网络信息检索，以及特定主题、媒体类型资源检索角度，探讨了不同任务难度下检索行为的特点和规律，以及基于行为的任务难度预测工作。这些研究大都从查询、浏览、点击、保存等方面选取行为指标，也有部分研究结合了检索策略[14]、搜索路径[15]等特征，还有研究者从时间阶段角度综合考虑行为特征[11]。这些研究往往通过实验室研究的方式采集数据，对行为数据和难度评价数据进行关联分析。可以看到，不同研究取得的发现并不完全一致，如在C. Liu等的两次研究中[11-12]，均发现用户在简单任务中查看的内容页面数量高于困难任务；而在J. Arguello[9]、J. Gwizdka和I. Spence[15]、J. Kim[16]的研究中则发现，用户在困难任务中会查看更多的网页。究竟哪些检索行为指标有助于更好地区分用户在不同任务难度感知下的状态？这会影响到后期对于基于行为进行任务难度预测的可靠性和稳定性。目前还较少有研究者进行多项实验间的比较工作，因此，本研究一方面结合3次不同的信息检索行为实验，另一方面也结合同类研究的发现进行综合横向比较，力图发现更为稳定的、对于任务难度感知有显著影响的行为指标，为后期开展预测提供参考和借鉴。

* 本文系国家自然科学基金项目“探索式检索过程中用户的情感体验及其影响机理研究”(项目编号:71974015)研究成果之一。
作者简介：黄崑，教授，博士，E-mail：huangkun@bnu.edu.cn；陈佳琦，硕士研究生；刘畅，副教授，博士。
收稿日期：2021-07-20 修回日期：2021-10-08 本文起止页码：83-92 本文责任编辑：王传清

2 研究设计

2.1 研究问题

本研究以学习型搜索为背景,结合现实的课程教学,开展检索行为实验,通过收集学生们完成搜索任务的行为数据和任务难度感知数据,分析显著影响难度感知的行为指标,并且与不同实验中的发现进行比较,评价指标稳定性。具体研究问题包括:①查询、点击、鼠标、查看、时间类检索行为特征,以及搜索策略特征在不同任务难度感知的搜索过程中,是否存在显著差异? ②在①中发现的显著性指标在不同的实验研究中,哪些行为指标在反映任务难度感知方面具有更好的稳定性?

2.2 检索行为指标

本研究在已有研究基础上[9-11],确立了查询、点击、鼠标、查看和时间 5 类检索行为,同时结合 R. Fidel[17]提出的检索策略类型,从操作型和概念型两个方面进行划分。其中,操作型策略是指通过调整检索系统功能来调整检索结果集合的操作;概念型策略是指通过修改检索提问的概念含义来调整检索结果集合的操作。此外,还将用户所浏览网页细分为 3 类:输入检索词前的"登录页";显示检索结果列表的"SERP 页面";显示单一检索结果摘要或全文的"Doc 页面"。因此,本研究构建的检索行为指标共计 7 类特征、50 项行为指标,分别是,查询特征(10 项)、点击特征(6 项)、鼠标特征(9 项)、查看特征(6 项)、时间特征(14 项)、操作型策略(1 项)以及概念型策略(4 项)。具体行为指标如表 1 所示:

表 1 检索行为指标体系

一级指标	二级指标	说明
查询特征	查询次数	输入检索词进行查询的次数
	查询总长度	所有查询提问长度总和
	查询平均长度	每次查询的平均提问长度
	首次查询和平均查询长度之差	首次查询长度与平均查询长度之差
	查询项总数	所有查询词总和
	查询平均项数	每次查询的平均查询词个数
	唯一查询项总数	所有非重复查询词总和
	查询项类型比	查询词重复率,即查询词总和/唯一查询词总和
	查询中停止词词数	查询中停止词总和
	查询中唯一停止词词数	查询中非重复停止词总和
点击特征	点击检索结果总数	在 SERP 页面点击检索结果的次数总数
	点击检索结果平均排名	点击检索结果在所有检索结果中的排名均值
	翻页点击总次数	在 SERP 页面点击前后翻页的次数总数
	平均翻页点击次数	每次查询在 SERP 页面点击前后翻页的次数平均值
	查询中点击次数	查询中没有点击操作标 0,有点击操作标 1
	查询点击百分比	有点击操作查询/总查询次数
鼠标特征	鼠标悬停总次数	鼠标指针进入了检索结果周围透明的浮动框中并有明显停顿的次数总数
	鼠标悬停平均次数	每次查询的鼠标悬停次数平均值
	鼠标悬停最高排名	悬停检索结果在所有检索结果中排名的最高值
	鼠标悬停平均排名	悬停检索结果在所有检索结果中排名的平均值
	SERP 页面鼠标下滑行为	在 SERP 页面的鼠标下滑行为总次数
	SERP 页面鼠标上滑行为	在 SERP 页面的鼠标上滑行为总次数
	Doc 页面鼠标下滑行为	在 Doc 页面的鼠标下滑行为总次数
	Doc 页面鼠标上滑行为	在 Doc 页面的鼠标上滑行为总次数
	鼠标移动距离(Pixels)	在任务过程中鼠标移动总距离
查看特征	查看所有 SERP 页面总数	查看 SERP 页面总数
	查看所有唯一 SERP 页面总数	查看非重复 SERP 页面总数
	查看所有 Doc 页面总数	查看 Doc 页面总数
	查看所有唯一 Doc 页面总数	查看非重复 Doc 页面总数
	查看所有网页总数	查看所有网页总数
	查看所有唯一网页总数	查看所有非重复网页总数

黄崑，陈佳琦，刘畅. 面向任务难度感知的检索行为研究[J]. 图书情报工作,2022,66(3):83－92.

（续表 1）

一级指标	二级指标	说明
时间特征	登录页平均停留时长	打开检索界面至输入检索词点击检索的平均停留时长
	登录页停留总时长	打开检索界面至输入检索词点击检索的总停留时长
	SERP 平均停留时长	SERP 页面的平均停留时长
	Doc 平均停留时长	Doc 页面的平均停留时长
	网页平均停留时长	网页的平均停留时长
	唯一 SERP 平均停留时长	非重复 SERP 页面的平均停留时长
	唯一 Doc 平均停留时长	非重复 Doc 页面的平均停留时长
	唯一网页平均停留时长	非重复网页的平均停留时长
	SERP 停留总时长	SERP 页面停留总时长
	Doc 停留总时长	Doc 页面停留总时长
	网页停留总时长	网页停留总时长
	文档编辑停留总时长	文档编辑总时长
	一次查询平均持续时长	每次查询的平均持续时间
	完成任务总时长	完成任务总时长
操作型策略	检索源调整	用户在检索过程中更换检索源的次数
概念型策略	在任务描述中出现有意义查询词数	在任务描述中出现的，在检索中使用的有意义查询词数
	没在任务描述中出现有意义查询词数	没有在任务描述中出现的，在检索中使用的有意义查询词数
	在任务描述中出现唯一有意义查询词数	在任务描述中出现的，在检索中使用的非重复有意义查询词数
	没在任务描述中出现唯一有意义查询词数	没有在任务描述中出现的，在检索中使用的非重复有意义查询词数

2.3 任务设计

本研究从全校通识教育课程《信息技术应用》中招募本科生，该课程主要面向低年级本科生进行 Access 数据库、VBA 程序设计相关知识的教学。笔者所在课题组（以下简称“课题组”）收集了往年多届学生在课程初期普遍关心的问题，并结合已有研究的任务设计，借鉴教育目标分类[18]和交互式检索任务认知复杂度特点[19]，设计理解、分析、评价、创造 4 类、5 项学习型搜索任务，具体任务描述如表 2 所示：

表 2　实验任务描述

类型	任务描述
理解类	Access 是怎样一个工具，最新版本是什么？最新版本与老版本相比增加了哪些新功能？请你在查阅相关信息后，给出说明
分析类	有同学不清楚如何在苹果操作系统下安装 Access，请你在查阅相关信息后，给他一些建议
评价类	国家计算机二级考试的科目包括 Access、C 语言、Python、Office 等，想知道考哪一科目对自己更有必要，为什么？请你在查询相关资料后，结合自身实际需求，谈谈你的看法
创造类	如果计划报考国家计算机二级考试 Access 科目，想了解考试内容有哪些，备考时间需要多久，如何制定备考计划？请你在查询相关资料后，做一个备考说明文档，保存为 WORD 文档
	本学期信息技术应用课程主要包括数据库原理、程序设计两大知识内容。因为感觉涉及很多数据，有些同学会担心自己数理成绩一般、日常计算机软硬件知识也不是很熟悉，担心这些方面薄弱的话会影响信息技术应用课程的学习。请你查阅相关资料后，谈谈你的看法，认为是否有影响？如果有影响，会产生什么影响？采用什么方法可以避免负面影响？如果没有影响，可以有哪些方法来提升学习效果？也请记录在文档中进行说明

2.4 数据收集

本次实验（即第 3 次实验）收集于 2020 年春季的《信息技术应用课程》，在第一次课程的时候以作业形式发放给学生，学生使用 Morae 软件记录信息检索、作业书写的完成过程。在完成作业前，使用 5 级李克特量表收集学生们对任务的难度感知、熟悉度、兴趣度等信息。每一项检索任务完成后，同样采用 5 级李克特量表收集学生对任务难度、困难原因及检索绩效（包括相关性、有用性、信心程度、成功感）的感知。这次实验中收集到 39 名学生提交的作业，男生 10 人，女生 29 人。其中，36 人来自政府管理学院，2 人来自经济与工商管理学院，1 人来自法学院。并且，36 人为大一新生，2 人为大二学生，1 人为大三学生。所有学生均使用个人电脑对搜索过程进行 Morae 录制，由于 Morae 软件卡顿、重启导致部分任务过程记录不完整，实际收集的有效检索行为数据为 170 条。

因为 Morae 录频问题，SERP 页面鼠标下滑行为和上滑行为、Doc 页面鼠标下滑行为和上滑行为、鼠标移

动距离、文档编辑停留总时长6项行为指标存在一定的缺失。但实验中行为变量缺失率最高为32.35%，低于缺失值阈值[20]，因此，根据多种缺失数据处理方法的对比分析[21]，本研究选取 A. P. Dempster 等[22]所提出 EM 算法，对因为录频软件问题导致的缺失行为进行变量填充。

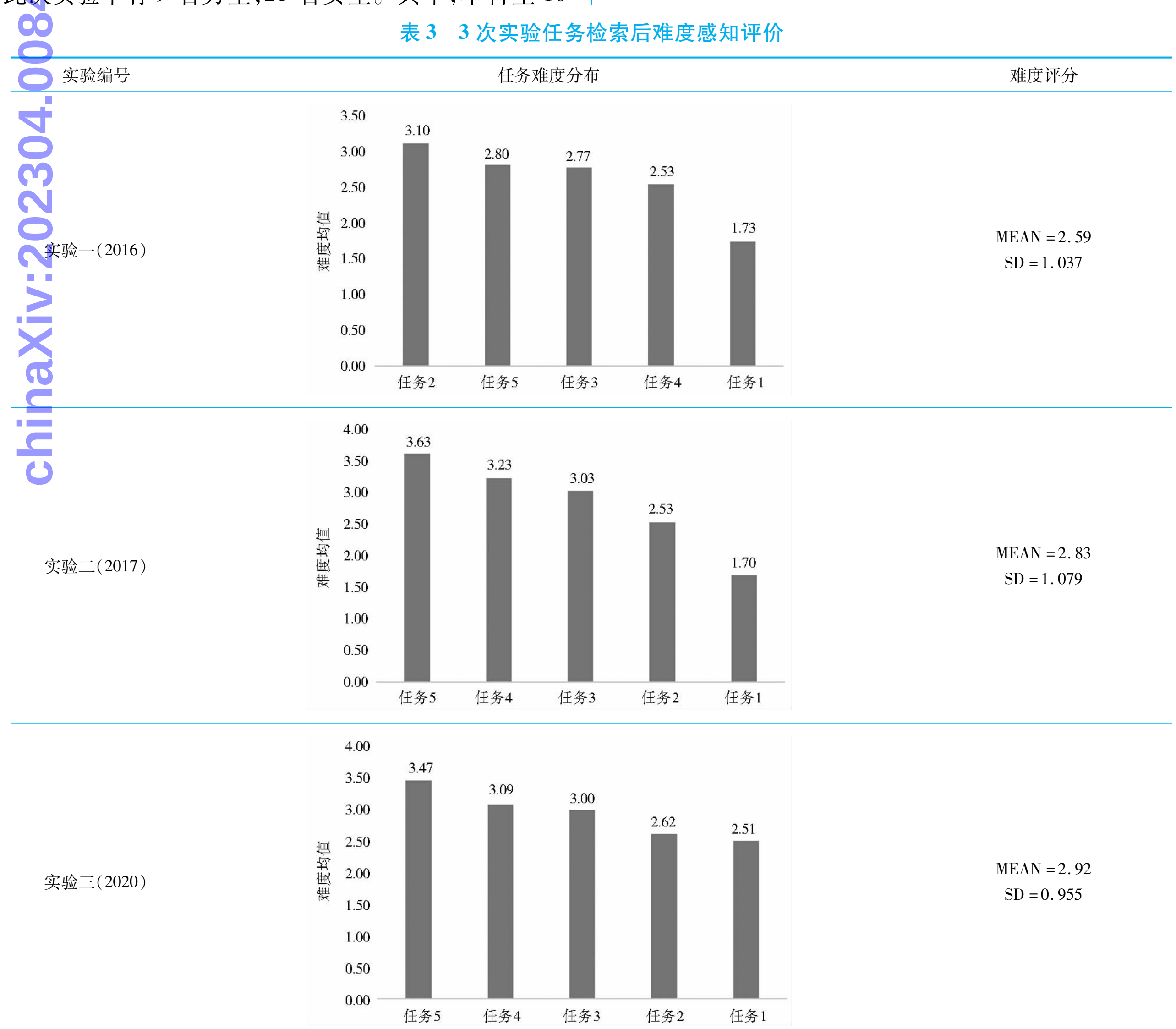
此外，本研究还综合运用了课题组前期的两次实验研究工作：①实验一[23]在2016年开展，以任务复杂程度[24]为依据，设计了事实、探索、抽象3类5项任务，随机招募30名在校学生进行实验室研究。被试中有男生5人，女生25人。其中，本科生8人，研究生22人。学生来自政府管理学院、地遥学院、生物科学学院等6个院系。②实验二[25]在2017年开展，根据教育目标分类法[18]，设计了记忆、理解、分析、评价、创造5类5项任务，随机招募30名在校学生开展实验室研究。此次实验中有9名男生，21名女生。其中，本科生16人，研究生14人，共来自21个学院。在这两次实验中，均收集了用户在任务完成前后对任务难度感知的评价，以及搜索过程的行为数据。本次研究利用3次实验数据，以及同类研究的发现进行横向比较分析，力图发现在不同实验任务、被试构成下，影响任务难度感知的更为稳定的行为指标。为使表述简洁，以"实验一"指代2016年开展研究，以"实验二"指代2017年开展研究，以"实验三"指代本文所开展最新研究。

3 数据分析与讨论

3.1 任务难度感知基本情况

以用户在检索后对任务困难程度的评分作为用户感知任务难度的判断依据，在总体情况方面，3次实验中的任务难度水平相当，3次实验任务难度分布及总体均值和标准差情况如表3所示：

表3 3次实验任务检索后难度感知评价



实验编号	任务难度分布	难度评分
实验一(2016)	难度均值：任务2 3.10；任务5 2.80；任务3 2.77；任务4 2.53；任务1 1.73	MEAN = 2.59 SD = 1.037
实验二(2017)	难度均值：任务5 3.63；任务4 3.23；任务3 3.03；任务2 2.53；任务1 1.70	MEAN = 2.83 SD = 1.079
实验三(2020)	难度均值：任务5 3.47；任务4 3.09；任务3 3.00；任务2 2.62；任务1 2.51	MEAN = 2.92 SD = 0.955

黄崑，陈佳琦，刘畅. 面向任务难度感知的检索行为研究[J]. 图书情报工作,2022,66(3):83－92.

3.2 显著性行为指标分析

在3次实验中选取检索后任务难度评价，将评分分值为4与5归为困难任务，3归为中等困难，1与2归为简单任务。3次实验数据均采用Kruskal-Wallis H检验方法，以比较在不同难度任务中，用户检索行为差异及其显著性。表4列出了3次不同实验中，在不同任务难度下呈现差异显著的行为指标数值情况。

表4　不同任务难度感知下差异显著的检索行为指标对比分析

行为指标	实验一（2016）				实验二（2017）				实验三（2020）			
	简单 N＝79	中等 N＝36	困难 N＝35	Sig.	简单 N＝56	中等 N＝50	困难 N＝44	Sig.	简单 N＝58	中等 N＝60	困难 N＝52	Sig.
查询次数	**2.75**	**4.08**	**7.31**	**0.00**	**4.11**	**6.54**	**6.25**	**0.00**	**2.69**	**4.12**	**4.42**	**0.00**
查询总长度	**13.39**	**16.58**	**26.97**	**0.00**	**18.00**	**39.36**	**36.03**	**0.00**	**20.16**	**39.88**	**39.44**	**0.00**
查询平均长度	4.80	3.94	3.63	0.44	**4.56**	**5.75**	**5.55**	**0.01**	**7.92**	**10.43**	**9.98**	**0.01**
查询项总数	**6.92**	**8.50**	**12.89**	**0.00**	**8.68**	**19.68**	**17.86**	**0.00**	**12.55**	**21.60**	**22.08**	**0.00**
唯一查询项总数	3.29	3.31	3.86	0.68	**4.50**	**7.14**	**6.30**	**0.00**	**7.97**	**10.70**	**10.92**	**0.00**
查询项类型比	**2.08**	**2.63**	**4.07**	**0.00**	**1.97**	**2.66**	**2.68**	**0.00**	**1.43**	**1.96**	**2.00**	**0.00**
点击检索结果总数	**1.62**	**3.31**	**3.94**	**0.00**	**4.45**	**6.36**	**5.05**	**0.02**	**3.16**	**5.35**	**4.73**	**0.00**
翻页点击总次数	0.43	0.81	0.80	0.34	1.36	1.52	1.64	0.50	**0.03**	**0.17**	**0.33**	**0.03**
平均翻页点击次数	0.13	0.20	0.17	0.47	0.34	0.34	0.39	0.54	**0.02**	**0.03**	**0.12**	**0.03**
查询中点击次数	**0.85**	**1.39**	**1.94**	**0.00**	**2.36**	**3.06**	**2.61**	**0.03**	**2.02**	**2.88**	**2.77**	**0.01**
查询点击百分比	33.26%	36.20%	29.68%	0.73	**62.98%**	**55.81%**	**47.52%**	**0.04**	**84.80%**	**73.91%**	**66.80%**	**0.01**
鼠标悬停总次数	－	－	－	－	**10.34**	**17.62**	**13.93**	**0.00**	**5.55**	**10.55**	**10.06**	**0.00**
鼠标悬停最高排名	－	－	－	－	**7.48**	**8.96**	**9.39**	**0.03**	**5.14**	**7.03**	**7.17**	**0.00**
鼠标悬停平均排名	－	－	－	－	－	－	－	－	**2.82**	**3.42**	**3.45**	**0.02**
SERP页面鼠标下滑行为	**3.24**	**6.33**	**7.37**	**0.00**	－	－	－	－	**52.26**	**147.72**	**72.31**	**0.00**
SERP页面鼠标上滑行为	**1.59**	**3.61**	**3.89**	**0.00**	－	－	－	－	**21.81**	**58.15**	**40.90**	**0.00**
Doc页面鼠标下滑行为	**0.94**	**1.67**	**3.31**	**0.00**	－	－	－	－	**193.66**	**570.47**	**267.37**	**0.00**
Doc页面鼠标上滑行为	**0.53**	**1.06**	**1.97**	**0.00**	－	－	－	－	**78.14**	**200.17**	**105.48**	**0.01**
查看所有SERP页面总数	**4.61**	**7.53**	**12.57**	**0.00**	**8.21**	**12.74**	**11.39**	**0.00**	**4.55**	**8.78**	**8.02**	**0.00**
查看所有唯一SERP页面总数	**3.19**	**5.03**	**8.31**	**0.00**	**5.04**	**7.56**	**7.30**	**0.00**	**2.91**	**4.52**	**4.98**	**0.00**
查看所有唯一Doc页面总数	**1.46**	**3.14**	**4.31**	**0.00**	**4.02**	**6.48**	**5.70**	**0.00**	**3.26**	**5.42**	**5.00**	**0.00**
查看所有网页总数	**6.24**	**11.17**	**18.00**	**0.00**	**13.07**	**20.90**	**18.68**	**0.00**	**12.41**	**19.12**	**17.98**	**0.00**
查看所有唯一网页总数	**4.65**	**8.17**	**12.63**	**0.00**	**9.05**	**14.04**	**13.00**	**0.00**	**6.17**	**9.93**	**9.98**	**0.00**
登录页停留总时长	**24.46**	**25.28**	**54.91**	**0.00**	35.91	46.52	55.48	0.18	**34.43**	**55.55**	**65.73**	**0.00**
SERP平均停留时长	18.86	18.93	14.54	0.21	19.21	19.87	20.67	0.18	**10.61**	**15.61**	**16.32**	**0.03**
唯一SERP平均停留时长	26.28	25.55	21.97	0.30	33.30	33.55	32.18	0.58	**18.61**	**30.33**	**27.84**	**0.01**
SERP停留总时长	**73.49**	**120.36**	**160.23**	**0.00**	**126.89**	**221.38**	**200.36**	**0.00**	**52.58**	**122.29**	**122.46**	**0.00**
网页停留总时长	**95.49**	**164.06**	**240.86**	**0.00**	**240.43**	**438.34**	**399.45**	**0.00**	**238.88**	**388.20**	**399.74**	**0.01**
文档编辑停留总时长	－	－	－	－	－	－	－	－	**253.43**	**354.20**	**368.50**	**0.03**
完成任务总时长	**95.49**	**164.06**	**240.86**	**0.00**	**240.43**	**438.34**	**399.45**	**0.00**	**563.95**	**818.01**	**898.40**	**0.00**
检索源使用个数	**0.46**	**0.78**	**2.51**	**0.00**	1.45	1.64	1.70	0.59	**1.22**	**1.30**	**1.50**	**0.05**
在任务描述中出现有意义查询词数	**5.85**	**5.94**	**9.40**	**0.05**	**6.25**	**11.10**	**11.98**	**0.00**	**8.36**	**14.98**	**15.04**	**0.00**
没在任务描述中出现有意义查询词数	**0.92**	**2.14**	**2.97**	**0.02**	**2.05**	**6.82**	**4.77**	**0.00**	**2.09**	**3.70**	**4.00**	**0.01**
在任务描述中出现唯一有意义查询词数	2.52	1.97	1.89	0.61	2.71	3.06	3.30	0.26	**4.88**	**6.30**	**6.29**	**0.00**

注：Sig≤0.05表明在0.05水平下差异显著，以加粗字体显示；实验一、实验二存在部分指标数值记录缺失，记为“－”

由3次实验数据分析可知：①在查询类的6项指标中，有4项指标在3次实验中都体现出了不同难度任务下的显著差异，均表现为高难度任务下行为次数显著高于低难度任务。②在点击类的5项指标中有2项指标，呈现显著差异，同样表现为高难度任务下行为次数显著高于低难度任务，不过对于中等难度和高难度，这两类任务之间的行为差异并不总是显著，但是简单任务与非简单任务之间的差异显著。③在鼠标类的

7 项指标中,前两次实验部分行为数据不完整,从实验一与实验三的比较,以及实验二和实验三的比较来看,尽管这些指标均呈现不同任务难度下的差异,但是在表现上并不完全相同,如中等困难的任务在“Doc 页面鼠标下滑行为”指标上显著高于简单、困难任务,并不完全遵循从简单到困难的递增变化规律。④在查看类的 5 项指标中,均在 3 次实验中体现了不同难度任务下的显著差异,简单任务下的行为次数要显著低于中等、困难任务。⑤在时间类的 7 项指标中,3 项指标体现出了不同难度任务下的显著差异,均表现为困难任务下持续时间长度显著高于简单任务。⑥在检索策略的 4 项指标中,2 个指标在 3 次实验中都体现出了不同难度任务下的显著差异,也表现为困难任务下使用的任务描述中有意义查询词的个数、未出现在任务描述中的有意义查询词的个数都要显著高于简单任务。

3.3　讨论

在经过上述比较,在 3 次不同实验中表现出不同任务难度感知下的显著差异的指标称为稳定性指标,其他均归为不稳定行为指标并进行剔除,最后获得 26 项稳定行为指标。这些指标意味着在不同的实验任务情境下具有反映用户难度感知差异的特点。同时,本研究还将 26 项指标与已有研究文献及发现进行横向比较,将任务难度感知对检索行为影响方向一致的指标称为共性指标,不一致的指标称为差异指标,未出现在已有研究中的指标称为新增指标。

为了分析检索行为指标与探索式搜索过程的关系,本研究借鉴 R. W. White 等[26]提出的浏览、集中检索两个基本过程,将检索行为归为探索式浏览和集中式检索两类,其中探索式浏览主要包含发现、学习与调查;集中检索主要包括提问词的再形成、结果分析等。如表 5 所示,26 项检索行为指标归为检索初期、检索过程两类,其次又将检索过程分为集中式检索与探索式浏览两类。

表 5　稳定行为指标体系

检索阶段		行为指标	相关文献	特征类型
检索初期		登录页停留总时长	新增指标	时间特征
检索过程	集中检索	查询次数	共性指标[9,11,27-29]	查询特征
		查询总长度	共性指标[27]	
		查询项总数	共性指标[9,11,28]	
		唯一查询项总数	共性指标[9,11]	
		查询项类型比	共性指标[9]	
		点击检索结果总数	共性指标[9,30-31]	点击特征
		翻页点击总次数	共性指标[9]	
		查询中是否有点击	共性指标[9]	
		查询点击百分比	共性指标[9]	
		检索源使用个数	新增指标	检索策略
		在任务描述中出现有意义查询词数	新增指标	
		没在任务描述中出现有意义查询词数	新增指标	
		在任务描述中出现唯一有意义查询词数	新增指标	
	探索浏览	鼠标悬停总次数	共性指标[9]	鼠标特征
		鼠标悬停最高排名	共性指标[9]	
		鼠标悬停平均排名	新增指标	
		查看所有 SERP 页面总数	共性指标[11]	查看特征
		查看所有唯一 SERP 页面总数	共性指标[11]	
		查看所有唯一 Doc 页面总数	差异指标[11]	
		查看所有网页总数	共性指标[16]	
		查看所有唯一网页总数	共性指标[15-16]	
		SERP 停留总时长	共性指标[11,29]	时间特征
		网页停留总时长	新增指标	
		文档编辑停留总时长	新增指标	
		完成任务总时长	共性指标[9,11,15-16,28,31-32]	

由表5可以看出：

(1)从任务开始初期用户对任务的解读和检索策略的构思过程来看，在困难任务中用户在检索登陆页停留时间(即从进入检索界面到输入检索词点击进行检索的时间)显著高于简单任务，反映出用户需要更多时间来酝酿和思考检索策略。C. C. Kuhlthau[33]曾在ISP模型中也指出，任务执行指出，用户对于检索目标都存在着模糊的不确定感。当用户感觉自己可能缺乏所需的领域知识、不太擅长时，会加大任务的难度感，并付出更多的努力去解决这个问题[34]。因此，在任务描述理解、联系已有的领域知识、选取关键词、构造检索式、选择信息源、预期检索结果等方面可能会耗费更长的时间。关于"登录页停留总时长"指标，在已有研究中有类似的指标，但不完全相同，本研究定义"登录页"为输入检索词前的检索界面，而J. Arguello[9]研究中定义"加载页"为点击检索结果后打开页面的时间，更类似于Doc页面浏览的一部分，所以在此将其看作是新增指标。

(2)从任务检索执行过程来看，用户会不断进行检索尝试，并且在查看、分析、对比检索结果中进行选择。这一过程中，多项研究都发现了在困难任务、简单任务之间具有显著性差异的行为指标。

首先，在集中检索方面，13项行为指标对于区分不同任务难度具有显著性，其中9项为共性指标，4项为新增指标。当用户应对更困难任务时，查询行为、点击行为次数显著高于简单任务。用户在困难任务下，会尝试对检索式不断进行调整、重构，使用不同的检索词也会更多，结果点击次数、翻页次数以及查询中产生点击的百分比都会显著高于在简单任务下的表现，表明用户会查看更多的信息，并且在查询过程中产生点击的情况也更频繁。类似的，A. Inthiran等[28]、J. Arguello[9]、J. Liu等[11]、A. Aula[27]、O. Dan[29]都曾发现，用户在应对更困难的任务时，查询次数显著高于更为简单任务的情况；A. Inthiran等[28]、J. Arguello[9]、J. Liu等[11]、A. Aula[27]表明，困难任务中的查询长度、查询词个数都要高于简单任务的情况。并且，用户在困难任务中表现为会尝试更多的检索调整。J. Kim的研究也发现，在困难任务中用户的检索调整次数显著提升[16]。随着任务复杂度的提升，检索调整的操作数也会随之增加[14]。

除了具体检索行为指标，本研究还从检索策略上新增了4项指标，并且均具有区分任务难度的作用。其中，操作型策略中主要体现为困难任务下，用户使用的信息源类型更多样，其中既包括不同的平台数，也包括在同一个平台的不同功能区。而对于概念型策略而言，用户在处理困难任务时，会更细致地从任务描述中抽取检索词，并且在搜索过程中受到启发去提出更多的、未在任务描述中提到的检索词。这些新的检索词既可能来自用户对结果的学习或发散，也可能来自系统推荐资源。X. Niu与D. Kelly[35]对用户使用具有查询建议功能的检索系统情况进行探究，结果表明，参与者将查询建议纳入其搜索速度非常快，对于专业知识经验较少的用户会使用更多的查询建议及保存更多文档。参与者在更困难的任务中，也会更多使用查询建议。通过查询扩展、查询词推荐、分面检索等方式，都有助于帮助用户获得更多启发，降低搜索寻找新词的困难。

其次，在探索浏览方面，12项行为指标对于区分不同任务难度具有显著性，其中8项共性指标，3项新增指标，1项差异指标。共性指标表明，当用户完成困难任务时，查看结果的最大深度，在可能有用的信息上的鼠标悬停次数、对所有SERP页面、唯一SERP页面的查看次数，以及网页总数和唯一网页总数都要显著高于简单任务，并且SERP停留时长、任务总时长也会显著高于简单任务。这些行为都是用户进行结果研读、筛选、结果相关性判定的基本活动。可见，在结果页面的更长时间阅读，查看更多不一样的结果，这些行为可能都预示着用户此时对任务存在困难感。类似地，J. Arguello[9]、X. Hu和N. Kando[31]、L. Fraser与C. Locatis[30]都曾发现，在困难任务中用户的鼠标点击次数显著增高；J. Arguello[9]研究中发现用户鼠标、滚动条操作在困难任务中操作次数更多；J. Arguello[9]、J. Liu等[11]、X. Hu和N. Kando[31]、D. Hienert等[32]、J. Gwizdka与I. Spence[15]、A. Inthiran等[28]、J. Kim[16]等均曾发现，用户在困难任务中花费的时间要明显多于简单任务。

在共性指标基础上，本研究新增3项指标。其中，"鼠标悬停平均排名"指标是综合前期研究优化所得，该指标的显著结果表明，在困难任务中用户会对检索结果列表进行更深度地浏览，而非仅仅浏览结果列表中排名靠前的检索结果。此外，本研究对SERP页面和Doc页面停留时间进行了区分，综合统计了所有网页的"网页停留总时长"，该指标的显著结果同样反映出用户在对搜索内容浏览、理解、学习方面所付出的努力。同时，"文档编辑停留总时长"是针对本研究开展的学习型搜索所新增的指标，用以反映用户在对信息

整合、汇总基础上完成任务目标所花费的认知努力，同样表现出在困难任务下，用户会耗费比简单任务更长的时间精力。

在探索浏览方面还有 1 项差异指标，即“查看所有唯一 Doc 页面总数”，J. Liu 等[11]发现“查看所有唯一 Doc 页面总数”随任务难度增加呈递减趋势。对比两次实验任务设计，本课题组所开展的实验，或者要求用户回答有关问题，或者要求其完成调研报告撰写，这对于查看 Doc 页面的需求更高，而 J. Arguello[9]和 J. Liu 等[11]的研究中仅要求用户保存或标记相关结果，这可能是导致实验结果差异的原因。

根据上述讨论可知，本研究通过多次实验比较，以及与同类研究的横向比较，总结了在不同研究条件下对于任务难度感知判断有显著作用的行为指标，有助于揭示任务难度感知与检索行为之间更为稳定的关系，为进一步量化映射及研究开展奠定基础，也能够为进一步任务属性预测研究的行为特征选取、权重设计提供参考。同时，结合探索式搜索过程的典型过程分析，为解释这些行为反映任务难度感知提供理论支持。

4 结语

本研究综合了文献调研和前期研究基础，依托课堂教学开展实验研究，以《信息技术应用》课程学生为研究对象，利用 Morae 软件通过远程方式开展检索行为实验，共收集到 39 名学生的 170 条有效检索行为数据。根据录屏数据和任务难度感知评价数据，对不同难度下呈现显著差异的行为指标进行分析。并且，结合课题组在 2016 年、2017 年开展的实验研究结果，以及与已有相关研究的发现进行综合比较，总结了 26 项对于区分任务难度感知有影响的显著性指标在不同研究中的结果。

在社会科学实验研究中，重复和再现相同的实验环境、条件面临许多困难和挑战，如实验招募的参与者存在主观性和个体差异，作为实验刺激的任务设计在引发参与者难度感知程度等方面，都存在精确衡量和控制的难度。但经过 3 次实验结果的比较可以看到，通过五级量表衡量任务主观难度感知，以及通过实验室统一机器进行实验和使用个人计算机进行实验，依然能够揭示出在不同实验任务设计下具有高稳定性的检索行为指标，其在不同任务难度感知下呈现显著差异，对于更稳定地进行任务难度感知预测奠定了基础。同时，未来研究还可以更深入地探讨影响指标稳定性的因素，探索不同特征组合方式，或结合时间维度，探索基于检索行为的难度预测建模策略和方法。并且，未来研究可结合具体检索情境特点，建立用户任务难度感知与多层次、不同颗粒度的检索行为之间的量化映射关系，实现基于行为的任务难度识别和判断功能，以嵌入信息系统中，使得检索系统能够在隐性的用户观察中，既可以洞察用户可能面临的搜索困境，及时提供相应的干预和调节，也可以优化任务分派机制，使用户在力所能及的情况下尽可能持续开展并完成搜索任务，避免在任务困难中退缩甚至停止。

本研究的不足主要是，虽然开展多轮实验进行检验，同时也与同类研究进行比较，揭示了在不同研究中任务难度感知与检索行为相对稳定的关系。但是，总体来看，实验研究的被试规模比较有限，所以在总体样本代表性上具有一定局限，还有待在更大规模的搜索情境下进行检验。同时，大多数实验室研究的任务设计来自研究主试布置，距离用户真实任务还有一定距离，所以后续可结合日志数据进行比对和印证，可以更好地检验本研究所提取的行为指标对于判断用户任务难度感知作用的普适性。此外，在行为指标方面，现有指标均为检索过程的累积结果，对于检索过程的实时难度感知反映不足，不同指标在区分困难感知程度上的差异也并未考虑，后续可对搜索过程中的任务难度感知及其变化进行更细粒度研究，以及对与检索行为相关的规律进行更深入地探索。同时，可以结合眼动、表情以及日志等更丰富的可观测数据，为任务难度分析提供支持，促进信息检索系统对用户搜索认知、情感状态的捕捉，提升用户检索绩效和体验。

参考文献：

[1] LIU J Q. Deconstructing search tasks in interactive information retrieval: a systematic review of task dimensions and predictors[EB/OL]. [2021-10-29]. https://doi.org/10.1016/j.ipm.2021.102522.

[2] LI Y L, BELKIN N J. A facted approach to conceptualizing tasks in information seeking[J]. Information process and management, 2008,44(6): 1822-1837.

[3] KIM S, SOERGEL D. Selecting and measuring task characteristics as independent variables[EB/OL]. [2021-10-29]. https://doi.org/10.1002/meet.14504201111.

[4] CRESCENZI A, CAPRA R, ARGUELLO J. Time pressure, user satisfaction and task difficulty[J]. Proceedings of the American Society for Information Science and Technology, 2014, 50(1): 1-4.

[5] 刘畅，宋筱璇，杨子傲. 用户信息搜索中的学习行为及过程探究[J]. 大学图书馆学报，2019，37(4)：36-45.

[6] 竺培梁. 心理测量——理论与应用[M]. 合肥：中国科学技术

大学出版社, 2008.

[7] LIU J J, LI Y, HASTINGS S K. Simplified scheme of search task difficulty reasons[J]. Journal of the Association for Information Science and Technology, 2019,70(5): 526 – 529.

[8] BELL D J, RUTHVEN I. Searcher's assessments of task complexity for Web searching[C]//Proceedings of the 26th European conference on information retrieval (ECIR 2004). Berlin: Springer, 2004: 57 – 71.

[9] ARGUELLO J. Predicting search task difficulty[C]//Proceedings of European conference on information retrieval. Cham: Springer, 2014: 88 – 99.

[10] LIU J J, GWIZDKA J, LIU C, et al. Predicting task difficulty for different task types[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1):1 – 10.

[11] LIU J J, LIU C, COLE M, et al. Exploring and predicting search task difficulty[C]//Proceedings of the 21st ACM international conference on information and knowledge management. Maui: ACM, 2012: 1313 – 1322.

[12] LIU C, LIU J J, BELKIN N J. Predicting search task difficulty at different search stages[C]//Proceedings of the 23rd ACM international conference on conference on information and knowledge management. Shanghai: ACM, 2014: 569 – 578.

[13] 黄崑, 陈佳琦, 郑明煊, 等. 信息搜索任务难度研究述评[J]. 信息资源管理学报, 2020, 10(4): 88 – 98.

[14] WILDEMUTH B M, KELLY D, BOETTCHER E, et al. Examining the impact of domain and cognitive complexity on query formulation and reformulation[J]. Information processing & management, 2018, 54(3):433 – 450.

[15] GWIZDKA J, SPENCE I. What can searching behavior tell us about the difficulty of information tasks? a study of Web navigation[J]. Proceedings of the American Society for Information Science and Technology, 2006, 43(1):1 – 22.

[16] KIM J. Task difficulty as a predictor and indicator of Web searching interaction[C]//Proceedings of CHI '06 extended abstracts on human factors in computing systems. Montréal: ACM, 2006: 959 – 964.

[17] FIDEL R. Moves in online searching[J]. Online review. 1985, 9(1): 61 – 74.

[18] ANDERSON L W, KRATHWOHL D R, AIRASIAN P W, et al. Taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives[J]. European legacy, 2001, 114(458): 1013 – 1014.

[19] KELLY D, ARGUELLO J, EDWARDS A, et al. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework[C]//Proceedings of the 2015 international conference on the theory of information retrieval. Northampton: ACM, 2015: 101 – 110.

[20] BARZI F, WOODWARD M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies[J]. American journal of epidemiology. 2004,160(1):34 – 45.

[21] 庞新生. 缺失数据处理方法的比较[J]. 统计与决策, 2010(24):152 – 155.

[22] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, series B(Methodological), 1977, 39(1): 1 – 38.

[23] 张路路,黄崑. 基于认知风格的数字图书馆用户信息检索行为研究[J]. 情报学报,2018,37(11):1164 – 1174.

[24] BYSTRÖM K, JÄRVELIN K. Task complexity affects information seeking and use [J]. Information processing & management, 1995, 31(2): 191 – 213.

[25] 王凯飞,黄崑. 数字图书馆用户检索失败的影响因素研究[J]. 图书情报工作,2019,63(11):25 – 34.

[26] WHITE R W, ROTH R A. Exploratory search: beyond the query-response paradigm[J]. Synthesis lectures on information concepts, retrieval, and services, 2009, 1(1): 1 – 98.

[27] AULA A, KHAN R M, GUAN Z. How does search behavior change as search becomes more difficult? [C]//Proceedings of the SIGCHI conference on human factors in computing systems. Atlanta: ACM, 2010: 35 – 44.

[28] INTHIRAN A, ALHASHMI S M, AHMED P K. A user study on the information search behaviour of medical students[J]. Malaysian journal of library & information science, 2015, 20(1): 61 – 77.

[29] DAN O, DAVISON B D. Measuring and predicting search engine users' satisfaction[J]. ACM computing surveys, 2016, 49(1):1 – 35.

[30] FRASER L, LOCATIS C. Effects of link annotations on search performance in layered and unlayered hierarchically organized information spaces[J]. Journal of the American Society for Information Science and Technology, 2001, 52(14): 1255 – 1261.

[31] HU X, KANDO N. Task complexity and difficulty in music information retrieval[J]. Journal of the Association for Information Science and Technology, 2017, 68(7): 1711 – 1723.

[32] HIENERT D, MITSUI M, MAYR P, et al. The role of the task topic in Web search of different task types[C]//Proceedings of 3rd ACM SIGIR conference on human information interaction and retrieval (CHIIR). New Brunswick: ACM, 2018: 72 – 81.

[33] KUHLTHAU C C. Seeking meaning: a process approach to library and information services [M]. New York: Libraries Unltd Inc, 2004.

[34] DERVIN B. Sense-making theory and practice: an overview of user interests in knowledge seeking and use[J]. Journal of knowledge management, 1998, 2(2): 36 – 46.

[35] NIU X, KELLY D. The use of query suggestions during information search[J]. Information processing & management, 2014, 50(1): 218 – 234.

作者贡献说明:
黄崑:确定研究选题,提出研究设计,全文统稿;
陈佳琦:参与研究设计,进行数据收集和分析,撰写初稿;
刘畅:参与研究设计与数据讨论分析。

Research on the Search Behaviors Oriented to Task Difficulty Perception

Huang Kun[1] Chen Jiaqi[1] Liu Chang[2]
[1] School of Government, Beijing Normal University, Beijing 100875
[2] Department of Information Management, Peking University, Beijing 100871

Abstract: [**Purpose/significance**] This study uses learning-related search as the background to explore interactive behavior indicators that have a stable and significant impact on users' perception of the task difficulty. [**Method/process**] This research relied on course teaching to carry out behavior experiments. It took Information Technology Application course students as the research objects, used Morae to record search behavior experiments remotely. A total of 170 valid search behavior records from 39 students have been collected. According to the screen recording data and the task difficulty perception evaluation data, this study analyzed the behavior indicators which show significant differences under different difficulty levels. In addition, this study also combined two experimental data that carried out by the research team in 2016 and 2017 and the findings of existing related research, horizontal comparison was made, so as to summarize the stability of significant differences of behavior indicators under different task difficulty levels. [**Result/conclusion**] Finally, a stable behavior indicator system with 26 behavior indicators is obtained. The research also describes users' behavior characteristics in difficult tasks from the perspective of exploratory search process and behavior types, and points out the future research directions.

Keywords: task difficulty search behavior interactive information retrieval experiment study